Anomaly Detection on Live Water Pressure Data Stream

Gal Petkovšek
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
gal.petkovsek@ijs.si

Matic Erznožnik
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
matic.erznoznik@ijs.si

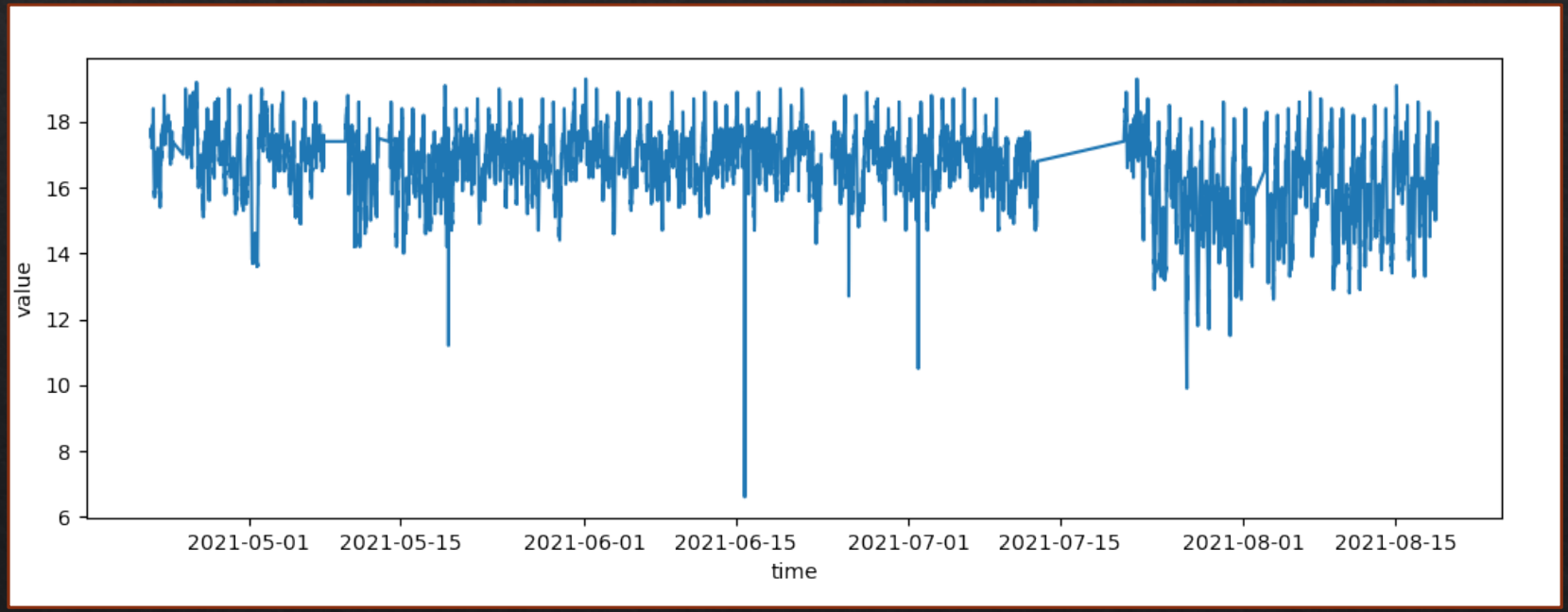
Klemen Kenda
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana
Slovenia
klemen.kenda@ijs.si

INTRODUCTION

- generic anomaly detection on live data streams
- we demonstrate already established anomaly detection approaches
- applied on pressure data of water supply system (Braila, Romania)
- the goal is to identify leaks
- unlabelled data
- evaluate algorithms using agreement rates

DATA

- 4 data streams from 4 separate sensors in the water supply network
- each stream contains ~10 000 samples
- 15 min intervals -> 100 days of data
- streams are split into a training (2000 samples) and evaluation part



EVALUATION OF ALGORITHMS

- evaluation on unlabeled data
- agreement rates

$$a_{\{i,j\}} = \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}\{f_i(X_s) = f_j(X_s)\}$$

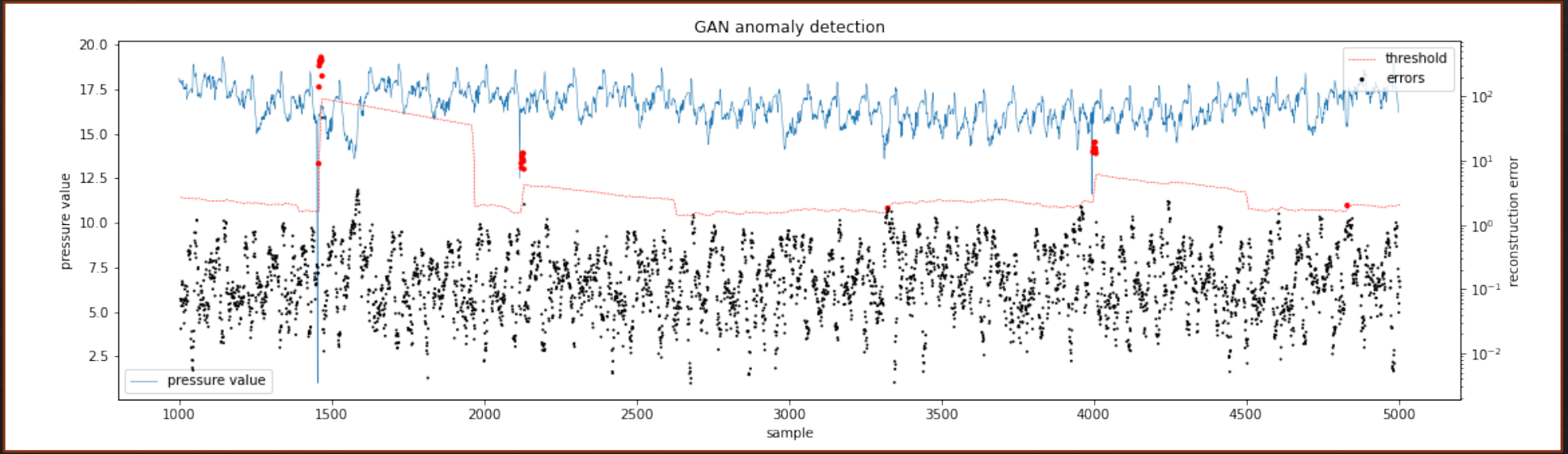
- estimating error rates

$$a\{i,j\} = 1 - e_{\{i\}} - e_{\{j\}} + 2e_{\{i,j\}}$$

- assumptions:
  - outperforms random classifiers
  - independent errors

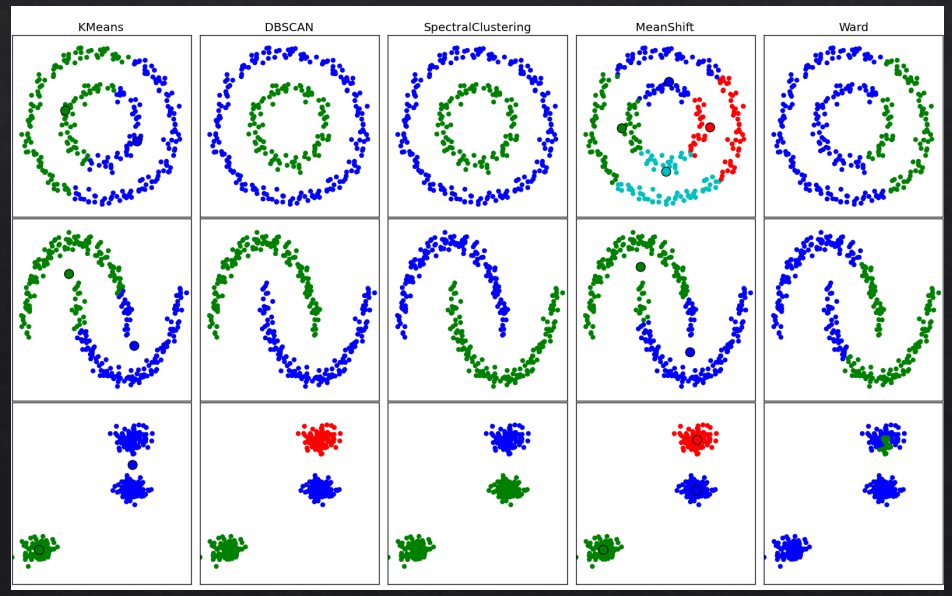
GAN

- Generative Adversarial Network
- encoder-decoder structure
- classify data points based on reconstruction errors
- the feature vector is composed of 10 consecutive pressure values
- we use a slightly modified threshold estimation method



DBSCAN

- clustering algorithm (Density-based spatial clustering of applications with noise)
- similar feature vectors to those used with GAN
- batch approach
- the largest cluster – normal; smaller clusters - anomalous



WELFORD'S ALGORITHM

- online mean and variance calculation

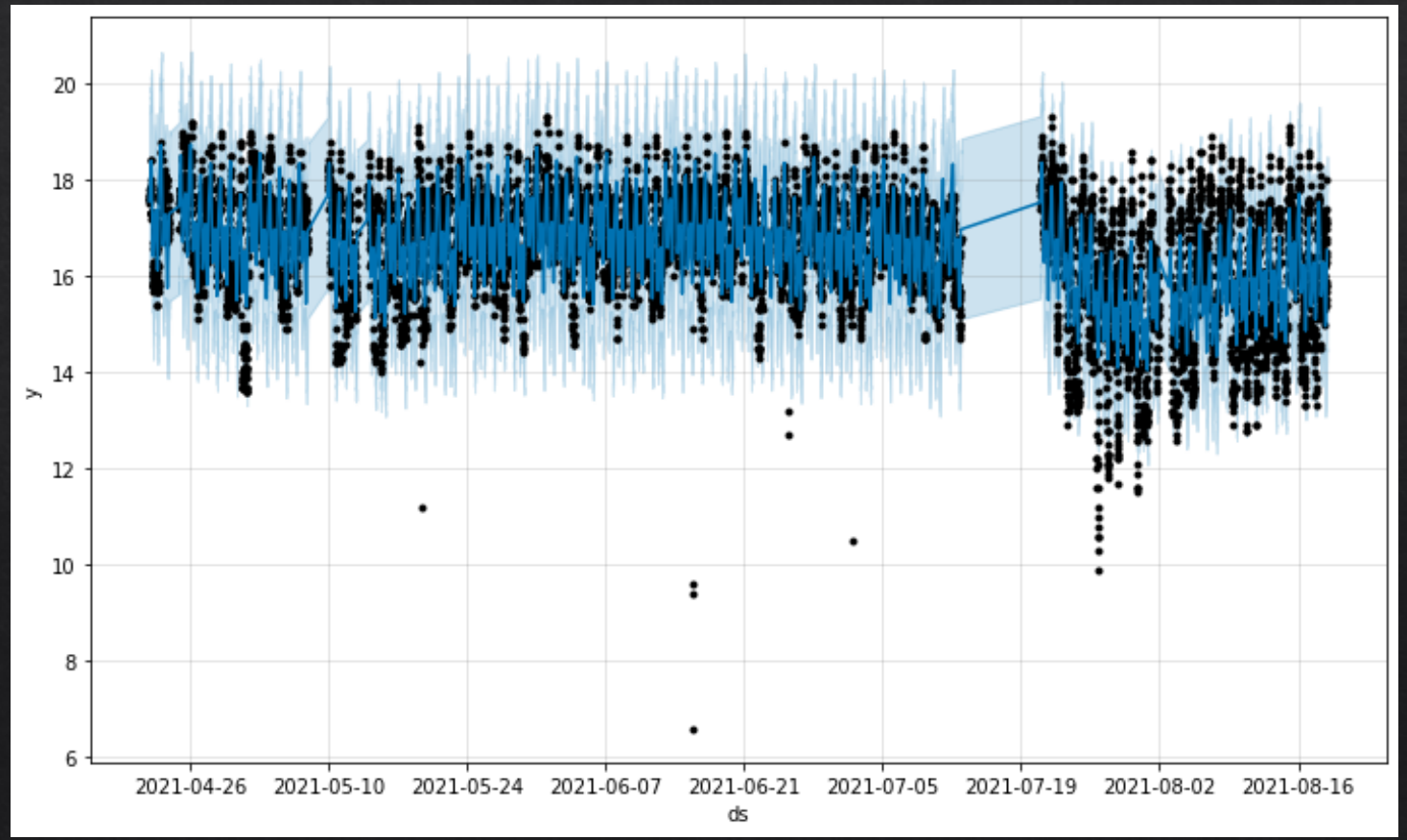
$$UL = mean + X * variance$$

$$LL = mean - X * variance$$

- Threshold
- Advantages: fast and simple

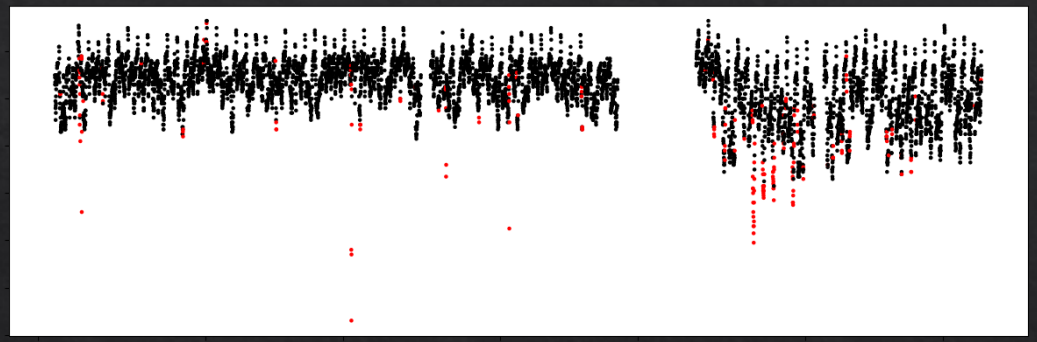
FACEBOOK PROPHET

- advantages:
  - sesonalities
  - missing data
  - easy to use
- confidence interval

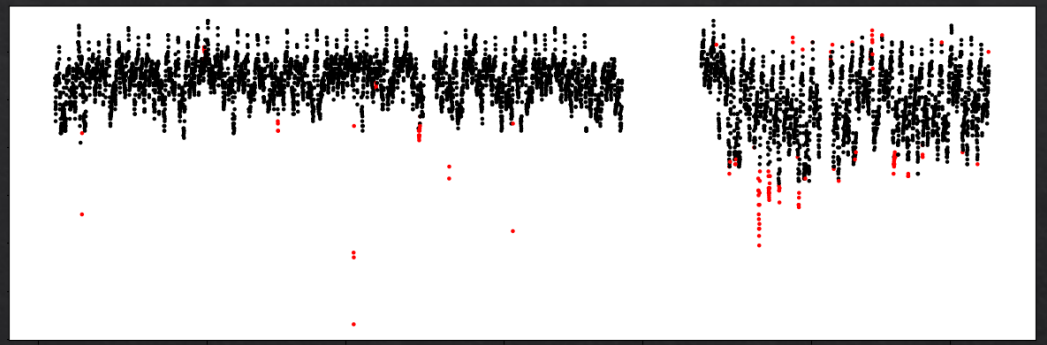


RESULTS

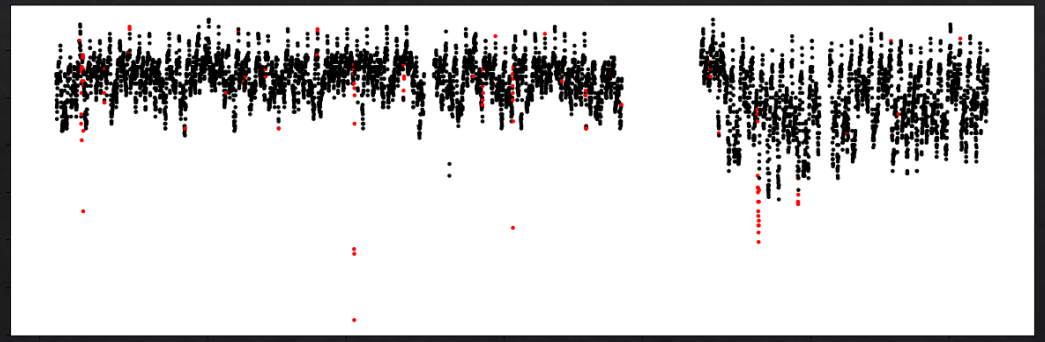
- visual comparison



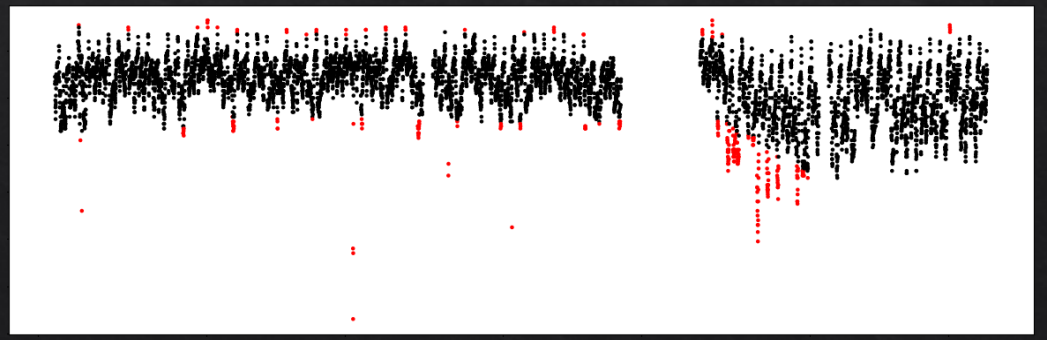
DBSCAN



Prophet



GAN



Welford's algorithm

- comparison of estimated errors

Algorithm	5770 Error rate	5771 Error rate	5772 Error rate	5773 Error rate
GAN	1.34%	1.38%	0.66%	1.09%
DBSCAN	1.59%	1.70%	1.78%	1.81%
Welford's algorithm	2.44%	2.41%	1.10%	2.31%
Facebook Prophet	1.14%	0.62%	0.39%	0.81%

- online setting
- isolation forest

CONCLUSION AND FUTURE WORK

- five algorithms, four data streams
- best performance
- best for online setting

- Facebook prophet for online anomaly detection